Identifying Invasive Ductal Carcinoma with Computer Vision

Siddharth Nanda, Vismita Uppalli, Dale Wilson
CS 4501: Computer Vision, University of Virginia
[sn9dq, vu8hh, dsw6ru@virginia.edu]

Abstract

Disease identification has been a hot topic in computer vision in the past few years as deep learning has become more feasible and gained traction. In our project, we explore the use of ML classifiers and pre-trained deep neural networks, including ResNet variants, in order to identify invasive ductal carcinoma (the most common type of breast cancer) from histopathology images. We utilized a Kaggle dataset created from sampling 277,524 50x50 pixel images from a set of 162 whole mount slide images of Breast Cancer scanned at 40x magnification in order to train (using a 90/10 train/validation split) k-nearest neighbors and softmax classifiers which were used to classify whether images contained instances of invasive ductal carcinoma. In addition, we utilized pre-trained variants of the ResNet neural network for the same classification task and compared the validation loss and accuracy of the models.

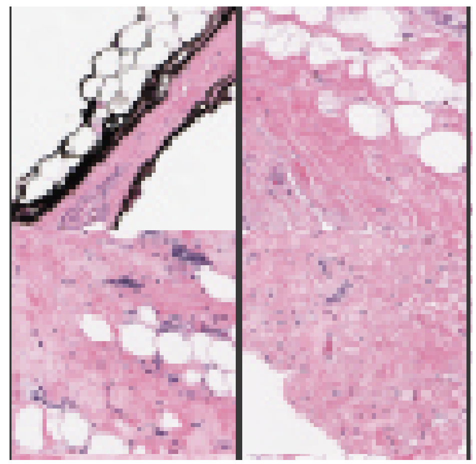

Figure 1. Here we show 4 of the sample images from Kaggle's IDC dataset. The images presumably include features medical professionals use to detect cancerous tissue (sarcomas, discoloration, etc.) We hope to train a model on these images that boasts an accuracy comparable to that of medical professionals

1. Introduction

Breast cancer is incredibly dangerous: it is the single most common type of cancer in women. One of eight women in the United States will develop breast cancer in their lifetime, and over 40000 women in the United States are expected to die from breast cancer in 2019 [2, 6]. Invasive Ductal Carcinoma (IDC) is the most common type of breast cancer, and about 80% of all breast cancers are IDCs [1] . The desire to quickly diagnose and treat diseases has spurred on a number of attempts to use newly developed deep learning technologies for that task. Several approaches attempt to make use of Artificial Neural Networks, computing models inspired by biological neural processes, since models such as ResNet have allowed researchers to utilize stunningly performant neural networks easily. In our work, we used histopathology images obtained from a publicly available dataset on Kaggle [5] and a variety of machine learning methodologies to attempt to accurately classify instances of invasive ductal carcinoma. We decided to implement various other models including a k-nearest neighbor and softmax classifier, as well as two pre-trained variants of ResNet as all of these models are quick to deploy with limited computing resources. Additionally, these models are easily customizable, allowing the ability to add layers and tune hyperparameters to the dataset. Breast cancer was chosen as the focus of this project as it is one of the few major cancers that when diagnosed early has an incredibly high survival rate (99%) and therefore is one of the diseases where using computer vision capabilities to quickly diagnose could be immediately impactful in terms of treatment.

2. Related Work

Work on identifying cancer with deep learning has focused on the use of convolutional neural networks (CNNs). In one study published this year, researchers developed a CNN to detect breast cancer from mammograms that only required lesion annotations during training in order to efficiently leverage the standard amount of medical annotations on a mammogram. They were able to achieve a remarkable per-image AUC of 0.88 on a dataset of digitized film mammograms [4]. Another study published this year focused on deployed convolutional neural networks for brain tumor segmentation: training and testing a CNN on a publicly available dataset of brain cancer images resulting in an accuracy of 75% [3]. Our work builds on the ideas introduced in both of these papers while integrating the performance benefits of utilizing pre-trained ResNet models.

3. Methodology

We used 3 different models to classify the Ductal Carcinoma images. The images used to train and test the models were resized to 256x256, cropped to 224x224, and normalized to parallel the nature of the images used to train and test the pre-trained ResNet classifiers. The first model used to classify the images is the Softmax Regression classifier. The model takes in flattened images of size 224 by 224 with 3 channels each. It consists of one linear layer which takes in a 224*224*3 flattened image vector and outputs a vector of size 2. The second model includes a K-Nearest Neighbor classifier which computes distances between the training and validation image sets. After finding the K nearest neighbors (evaluated at 1, 5, and 10), the most frequent category was determined as the classification for the image. The third model includes a pre-trained version of ResNet18 and ResNet50. The images were trained on the ResNet18 with 3 added linear layers. The first had an input of 1000 and output of 750. The second had an input of 750 and output of 250. The third had an input of 250 and output of 2. The same three linear layers were added to the ResNet50 as well. Models output predictions for two classes, cancer positive (+) and cancer negative (-).

4. Analysis of Results

We used pytorch to develop our three different models: Softmax, K-Nearest Neighbors, and ResNet classifiers. We first tested the Softmax Regression classifier which yielded a peak accuracy of 0.8573 (See Figure 2) after training for 15 epochs with a learning rate of .0001. Figure 2, for both training and testing, shows a steady training accuracy, slightly increasing with epoch, and a validation accuracy with little variation. In search of better performance, we implemented a K-Nearest Neighbors classifier. The KNN classifier yielded an accuracy of .8160, .8463, and .8502 when classifying based on 1, 5, and 10 nearest neighbors respectively (See Figure 3). After implementing the KNN classifier, we decided to train two ResNet classifiers using the pre-trained architectures of the ResNet18 and ResNet50 models. Both models were trained for 10 epochs with a learning rate of .01 because of the prohibitive nature of how long the ResNet models took to train. The sheer number of layers as well as the complexity of their architectures contributed to the lengthy train time of each model. The ResNet18 model had a peak accuracy of 0.9121 (See Figure 4) while the ResNet50 model had a peak accuracy of 0.9308 (See Figure 5) and consistently performed with an accuracy above .9000. Both of these figures show a decreasing loss with epoch, and increasing accuracy with slight variation for training and validation. We believe that the integration of such models directly into the histopathology imaging process would be a useful tool to allow doctors to see a diagnosis before exerting their clinical expertise.

5. Conclusion

Our models demonstrated acceptable performance for the classification task with a minimum accuracy of 0.8160 (KNN classifier with 1 nearest neighbor) and maximum accuracy of 0.9000 (ResNet). Both of these models were somewhat slow to test with limited computing resources, but cloud application with greater access to computing resources could be directly integrated with histopathology imaging software in practical use. We believe the use of deep learning based computer vision methodologies demonstrated in this paper could definitely be of use in a clinical environment: helping doctors receive a quick second opinion in their quest to accurately diagnose and treat diseases. However, such models must be: tested with far greater rigour on a much larger dataset, must be more performant, and must demonstrate less variance in accuracy.

References

[1] breastcancer.org. Invasive Ductal Carcinoma (IDC). `https://www.breastcancer.org/symptoms/types/idc`, 2019. [Online; accessed 10-December-2019].

[2] breastcancer.org. U.S. Breast Cancer Statistics. `https://www.breastcancer.org/symptoms/understand_bc/statistics`, 2019. [Online; accessed 10-December-2019].

[3] S. Iqbal, M. U. Ghani Khan, T. Saba, and Mehmoo. Deep learning to improve breast cancer detection on screening mammography. *Microscopy Research and Technique*, 82(8):1302–1315, 2019.

[4] J. H. R. E. F. R. M. . W. S. Li Shen, Laurie R. Margolies. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9, 2019.

[5] P. Mooney. Breast Histopathology Imagess. `https://www.kaggle.com/paultimothymooney/breast-histopathology-images`, 2017. [Online; accessed 10-December-2019].

[6] D. of Cancer Prevention, C. f. D. C. Control, and Prevention. Breast Cancer Statistics. `https://www.cdc.gov/cancer/breast/statistics/index.htm`, 2019. [Online; accessed 10-December-2019].

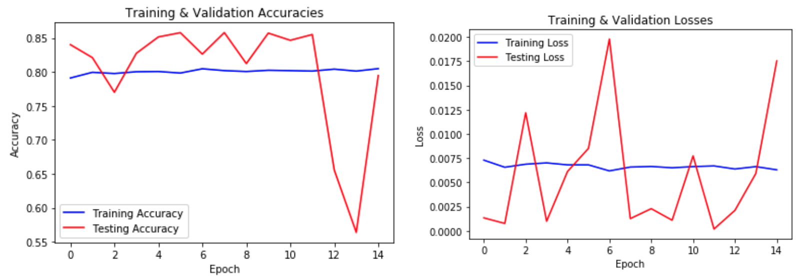

Figure 2. Here we show a figure detailing the training and validation accuracy and losses for the Softmax regression classifier

	K = 1	K = 5	K = 10
all	81.06%	84.63%	85.02%

Figure 3. Here we show a figure detailing the validation accuracies for the K-Nearest Neighbors classifier at K = 1, 5, and 10

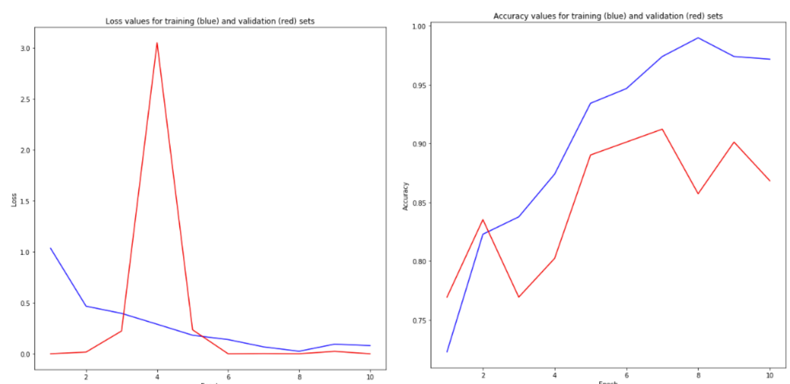

Figure 4. Here we show a figure detailing the training and validation accuracy and losses for the ResNet18 classifier

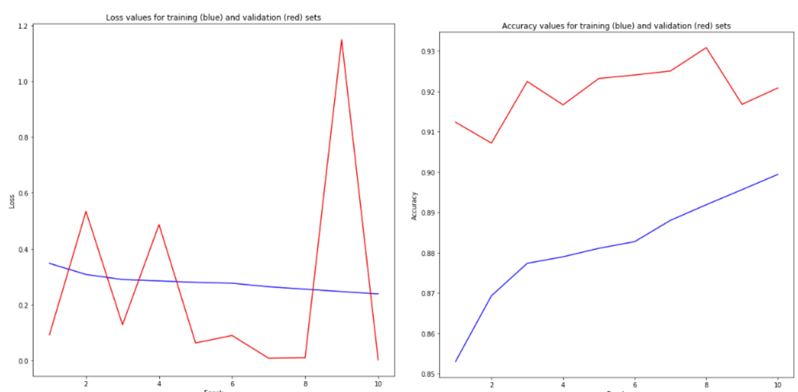

Figure 5. Here we show a figure detailing the training and validation accuracy and losses for the ResNet50 classifier